Architecture Workspace and V/CDE Workspace Liaison Meeting

Date, Time & Location:	5/18/04, 1-3 PM, NCICB, 6116 Executive Blvd. Rockville, MD (by teleconference)
Attendees:	Roll call was performed with the following participants attending: **Architecture Liaisons** Fred Hutchinson • Robert Robbins Ohio State • Scott Oster • Tahsin Kurc **V/CDE Liaisons** Mayo • Harold Solbrig UC-Davis • Cecil Lynch Albert Einstein • Xin Zheng NCI/OC/CIPS • Frank Hartel **Other Attendees** Washington University • Rakesh Nagarajan University of Pittsburgh • Jim Harrison University of Pittsburgh • Jim Harrison University of Hawaii • Leo Cheung Jackson Lab • Jim Kadin City of Hope • Joyce Niland • Hemant Shah • Jennifer Neat OHSU • Laura Fournier EMMES Corporation • Claudia Valmonte • Ryan Campbell NCICB • Peter Covitz • Leslie Derr • John Qu • Larry Wright • Margaret Haber Fred Hutchinson

	• Dan Geraghty Jackson Lab • James Kadin Coldspring Harbor • Michael Townsend University of Wisconsin • Rhoda Arzoomanian SAIC • Kathleen Gundry • Juerten Lorenz BAH • Arumani Manisundaram • Christine Richardson • Mike Keller
Agenda Item #1:	**Goal of Meeting/Introduction** Peter Covitz opened the meeting by stating that the Architecture/VCDE group needs define caBIG 'compatibility', and what it means for caBIG Workspace products or artifacts to be caBIG-compliant. This does not mean drafting scenarios, but providing guidance across caBIG community. • The two Cross-cutting Workspaces need to be working together at all times. • The Architecture WS has decided to break into sub-groups, one of which is Information Architecture. They are involved with determining how the data is portrayed in the space as well as re-distributing the meaning of that data in a grid-type fashion. This corresponds with issues surrounding domain models and data representation. • It is possible to get some basic recommendations together for developers pretty quickly depending on the outcome of this meeting. We will spend a fair amount of time on Agenda Item #2. • A broad caBIG requirement might be characterized as, "I sit down in front down of my caBIG console" and I want to: o Find what data are available o Determine what they mean o Find out how they are represented o Determine how the data fit into the broader space of biomedical information • The goal of this meeting is to determine how the answers to these questions are going to be formalized. • At model level, we use UML (modeling language-formal way of describing entities or classes). UML is accessible; additionally it is a formalism and can be fed into other software tools. • UML is great for describing a broad domain of interest. But we need, also, to get down to nuts and bolts; UML does not allow us to go another level of granularity. CDEs are little pieces represented

	in UML. The next level down in granularity are the terms (or values) that populate the data with (controlled vocabulary or ontology)
	• Essentially, caCORE has broken these down into 3 chunks with different levels of granularity:
	  o UML
	  o CDE
	  o Terms to define data and semantic standards (Vocab)
	• **What is your feedback?**

Agenda Item #2:	**Metadata and Domain Models** Harold Solbrig introduced the SAGE Project at Mayo. This project is trying to create interoperable guidelines to be able to reference data in a neutral fashion. • The terminology was the key set of definitions throughout the model. The terminology was key to coming up with attributes and possible values for the attributes. • Found it was necessary to anchor the terminology all the way up the spectrum. • An organized set of definitions needs to permeate the information model. Bob Robbins briefly gave his background emphasizing that his is a basic science and not a clinical background. His experience tells him that if caBIG is to succeed then the following items need to be addressed: • 'Meaning' change over time • Scientists can agree on the term, but multiple meanings or definitions evolve • Need some sort of universal naming authority • Interdatabase referential integrity. No one has built infrastructure like a cascade or a notification that could effect a foreign key in another database Cecil Lynch next brought up the topic of semantic drift. • Cecil Lynch: A term that has experienced extensive semantic drift needs to be handled as a different (or new) term. • Cecil Lynch: Doesn't see how multiple definitions would work out • Peter Covitz: caBIG will have a forum to reconcile these issues • Semantic drift, proposed multiple definitions, way to handle it if there is that much semantic drift then it has to be a different definition • Cecil Lynch: One example is with the words gene and locus. These are basic concepts that have acquired slightly different meanings to different people • Harold Solbrig: Ontologies and terminologies are never nicely

partitioned, we need to manage some fuzziness, what we want to do as much as possible is agree with what the thing is

- Peter Covitz: Permeation of the information system with vocabulary is key. 5 years go by, you are retrieving the operational definition that was there when the data was collected and then you determine if that definition is still valid.
- Harold Solbrig: If you put a definition on a data element, it is quite possible that one thing might be identified in many ways. Publish standard naming mechanism. That's where you need shared information model

Cecil Lynch returned to the conversation to the topic of referential integrity.

- Cecil Lynch: It seems to me that you would never be able to maintain concurrency across disparate databases with different key structures. A solution to this would be to look to the metadata registry
- Harold Solbrig: At bare minimum, it is crucial to just come up with common names. Next step is to be able to publish or make available info that others need or find useful.
- Rob Robbins: Referential integrity foreign key vs. primary key
- Peter Covitz: Make an assertion we do not want to put direct access, this is behind grid interoperability. This problem of naming is everywhere on the web (Universal Reference identifiers is something we could potentially use that)
- Rob Robbins: Another problem is data connectivity degrades. Most commonly things are either lumped of split. For example one gene is actually 2 genes with an intervening region or a common cause is determined for two diseases and they are then lumped together. How do you deal with this?
- Frank Hartel: SNOWMED and NCI thesaurus keeps a history of a term life, the entire life cycle of every concept in vocabulary is kept and can be used to determine what the current terminology is and what all the predecessors were, dated. Available both at concept and term level with SNOWMED, can be leveraged

ACTION ITEM: Develop a series of guiding principles and ask if they are caBIG compliant. Begin with caCORE and determine if it is sufficient and go from there.

- Joyce Niland: Is mapping of CDEs to SNOWMED and LOINC a guiding principle? Should new terminology be created only if the CDE does not exist?
- Peter Covitz: A gap in caCORE is that vocabulary does not permeate caCORE.
- Peter Covitz: EDRN system (from Hutch) uses CDEs for semantic continuity.
- Peter Covitz: EDRN uses CDEs to achieve interoperability. However, it does not provide any real formal modeling environment to establish relationships between data elements. That's where UML comes in, allows you higher order relationships across data

elements.

- Peter Covitz: EDRN is used as an example and discussion. IS the CDE sufficient to define interoperability? Do we need models, such as classes? For example, not just a list of genes, but genes in a certain pathway?
- Harold Solbrig: Having looked at CDEs, the higher level grouping exist whether formalized or if just put in the name. The left hand side of the name is starting to be the class.
- Harold Solbrig: It is not totally obvious of how formal we have to go. Advantage of UML, when you put things together graphically, it helps clarify some bad misunderstanding which can occur. Would advocate for representing model like relationships across data elements.
- Cecil Lynch: I would agree completely. Another point, the issue of when a term does not exist in controlled vocabulary, the idea of creating a new term versus adding a new term.
- Cecil Lynch: NCI has anatomical terms; there are terms in SNOWMED that are not in NCI, so add the terms from SNOWMED to NCI rather than creating them.
- Harold Solbrig: There is not a clear boundary between the information model and the vocabulary mode. There are things that need to be clarified.
- Peter Covitz: Do you think we need an identifier system that goes all the way down to the data object? Or is it sufficient instead for having a universal identifier system to define a class? We say what a gene is, but not specific genes. Is that a tolerable amount or not?
- Rob Robbins: Want decent reliability in quality of data objects in caBIG if someone is going to go on the grid and use what is on there for some computational analysis.
- Rob Robbins: Thinking about a way to come up with unique identifiers for caBIG sources. Make sure that the source and the object are unique
- Harold Solbrig: With gene identifiers you are at the mercy of what is the accepted practice. Need to cope with multiple identification schemes.
- Jim Kadin: You have to have identifiers of the individual genes; the only way to refer to the gene is by an identifier. Gene, sequences, clones, snips have to have identifier.
- Jim Kadin: Multi-part identifiers once you start applying attributes. You don't want to change identifier if the object type changed. Ultimately have to have identifiers for a lot of things. Adopt identifiers that already exist.
- Frank Hartel: We name genes consistently from outside authorities, when UNIPROT comes along we will look to that for identifiers
- Peter Covitz: Starting to hear an acceptance that at least for some types of well established models being used as identifiers in a new

database

- Rob Robbins: I disagree. Only submitters can update submissions, other databases allow anyone to update submissions. Maybe put caBIG in front of an identifier
- Cecil Lynch: How do you handle overlap?
- Peter Covitz: Hierarchal identifying system, resolution of data. How far do we want to go on rules for Object Permanence? Maybe classes of data where immutability of naming is important.
- Peter Covitz: Want to return to the topic of vocabulary mapping through the information model. If we have classes, and attributes of those classes, and instances of those classes (which are the data objects), what do we need to provide a caBIG compatible grid service? Every data class, as well as every attribute, has to have a definition.
- Harold Solbrig: Not necessarily have a well-specified definition, but reference an appropriate definition system.
- Peter Covitz: What level do you need to define the values themselves? Do we demand this? With CDE there is a field to define your values. In caDSR world, for example, M and 1 would both mean male; F and 2 would me female. Unique meaning…even though used differently, still has unique meaning. .
- Margaret Haber: LOINC mapping is in the NCI metathesaurus . The UMLS includes LOINC so the base code in in UMLS.
- Harold Solbrig: Is it your intent to assign a NCI code to every LOINC code?
- Margaret Haber: No. We end up with NCI code when things aren't represented in that UML code
- Harold Solbrig: In Sage/HL7, when you use it, 90% of what you need is there.
- Margaret Haber: We can refine the mappings when you load the local terms.

Peter Covitz then provided a summary of the conversation resulting in these main points:

- Vocabulary needs to permeate info model
- Strategy for concept history and guarding against semantic drift
- Need to have universal identifier system (combine source and object identifier, suggested to be consistent with HL7)
- Definitions must accompany data (do not define in your mind)
- Standardized naming conventions for classes
- Some type of higher modeling representation (UML or something comparable) that allows you to graphically link data classes
- Multiple definitions of same term need to be managed
- Need mechanism to define translation services from local to shared

Agenda Item #2:	**Management of Vocabulary and CDE sources**

The following questions were raised by Peter Covitz:

- Do we create repository for people to deposit their models?
- Do we use using caDSR?
- Do we using EVS?
- Mayo charged with next generation vocab server?
- Do we envision a federation of these services?
- What is practical for caBIG to deploy?
- Xin Zheng: Global services managed by global authority, then have sub-group authority
- Harold Solbrig: To a degree, the answer is to work through more use cases
- Peter Covitz: Whatever system generates those identifiers, there has to be some central identifier.
- Harold Solbrig: One form of identifier system is delegation
- Peter Covitz: Deployment topology is still a little early, need to know more
- Harold Solbrig: both use case and volume
- Peter Covitz: Arumani/Christine get notes together and distribute and have informal discussion with liaisons to determine action items and what deliverables such as white papers and high level recommendations (areas of further specification) will come from this.

Other discussion items:

Action Items:

Name Responsible	Action Item	Date Due	Notes
Christine/Arumani	Discussion with Liaisons & determine action items.	May	